W54–W58 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 11 June 2012*
doi:10.1093/nar/gks572

VARIANT: Command Line, Web service and Web interface for fast and accurate functional characterization of variants found by Next-Generation Sequencing

Ignacio Medina[1], Alejandro De Maria[1], Marta Bleda[1,2], Francisco Salavert[1,2], Roberto Alonso[1], Cristina Y. Gonzalez[1] and Joaquin Dopazo[1,2,3,*]

[1]Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), [2]CIBER de Enfermedades Raras (CIBERER) and [3]Functional Genomics Node (INB) at CIPF, Eduardo Primo Yufera 3, Valencia, 46012, Spain

Received February 19, 2012; Revised May 18, 2012; Accepted May 21, 2012

ABSTRACT

The massive use of Next-Generation Sequencing (NGS) technologies is uncovering an unexpected amount of variability. The functional characterization of such variability, particularly in the most common form of variation found, the Single Nucleotide Variants (SNVs), has become a priority that needs to be addressed in a systematic way. VARIANT (VARIant ANalyis Tool) reports information on the variants found that include consequence type and annotations taken from different databases and repositories (SNPs and variants from dbSNP and 1000 genomes, and disease-related variants from the Genome-Wide Association Study (GWAS) catalog, Online Mendelian Inheritance in Man (OMIM), Catalog of Somatic Mutations in Cancer (COSMIC) mutations, etc). VARIANT also produces a rich variety of annotations that include information on the regulatory (transcription factor or miRNA-binding sites, etc.) or structural roles, or on the selective pressures on the sites affected by the variation. This information allows extending the conventional reports beyond the coding regions and expands the knowledge on the contribution of non-coding or synonymous variants to the phenotype studied. Contrarily to other tools, VARIANT uses a remote database and operates through efficient RESTful Web Services that optimize search and transaction operations. In this way, local problems of installation, update or disk size limitations are overcome without the need of sacrifice speed (thousands of variants are processed per minute). VARIANT is available at: http://variant.bioinfo.cipf.es.

INTRODUCTION

Exome or genome sequencing constitutes a promising instrument for finding novel mutations of human disorders (1,2). However, massive sequencing experiments reveal an enormous amount of genomic variations (3), mostly of unknown functional consequences. Moreover, it has been noted that most sequence variants found in patients are likely to be neutral and do not cause any severe disorders (2). Within this scenario, the identification of casual mutations still represents a big challenge. Typically, several filters are applied to reduce the number of candidate variants, which include a crucial step of functional assessment.

Software analysis pipelines currently used in the analysis of Next-Generation Sequencing (NGS) data are highly modular, heterogeneous and rapidly evolving. Apart from several commercial packages, there are also open source packages available such as SAMTOOLS (4), GATK (5), GAMES (6) and others. Some of them provide information on the consequence type of the variants found. However, the limited nature of this information has fostered the recent development of tools specifically designed for the annotation and functional assessment of Single Nucleotide Variants (SNVs) such as ANNOVAR (7), snpEff (http://snpeff.sourceforge.net/) or Variant Effect Predictor (8). VARIANT can be considered as a new generation tool for the functional assessment of variants because: (i) it extends its information contents from coding to non-coding regions too, and (ii) it implements technological novelties such as RESTful Web Services that optimize search and transaction operations and allow

*To whom correspondence should be addressed. Tel: +34 96 328 96 80; Fax: +34 96 328 97 01; Email: jdopazo@cipf.es

using a remote database in an extremely efficient way. VARIANT has three clients: Command Line Interface (CLI), Web Application and Google Chrome Extension.

SUMMARY OF THE FEATURES OF VARIANT

VARIANT can easily be incorporated into a NGS-resequencing pipeline either as a CLI or invoked as a Web service. In addition, it can be invoked through a web application as a conventional web server. VARIANT inputs data directly in Variant Call Format (VCF) (9), which is the output of the most widely used programs for variant calling. VARIANT can report the functional properties of any variant in all the human, mouse, rat, zebrafish or fruitfly genes (and soon new model organisms will be added). VARIANT not only reports the obvious functional effects in the coding regions but also analyses SNVs in non-coding regions situated both within the gene and in the neighborhood that could affect different regulatory motifs, splicing signals and other structural elements or evolutionarily highly conserved elements. In addition, known phenotypic or disease-related variants from the GWAS catalog, OMIM, COSMIC mutations, etc., are reported.

Biological features

VARIANT reports the conventional consequence type of the variant that can be: Non-synonymous, Synonymous, Intronic, 5′ UTR/3′ UTR, Upstream/Downstream, Essential Splice Site, Splice site, Stop gained, Stop lost and Intergenic and Non-coding and Nonsense-mediated decay transcripts. We use as a reference consequence type the Sequence Ontology (10) from Open Biological and Biomedical Ontologies (OBO) (11), together with Ensembl consequence types and National Center for Biotechnology Information (NCBI) terms.

In addition, VARIANT reports information on variants that affect different regulatory or structural sites, such as CCCTF-binding factor (CTCF) transcriptional repressor sites, polymerase and histone sites or open chromatin regions, taken from Ensembl (12). Also variants disrupting transcription factor-binding sites (TFBSs) from the Jaspar database (13) or miRNA targets (14) are reported. Highly conserved regions between human and mouse genomes, taken from the Ensembl (12), are also reported because of their putative functional relevance. Another useful information to decide on the possible functional effect of an already described variant is the HapMap (15) allele frequency, which is also reported by the program. VARIANT also provides calculations of selective pressure values, related to the functional impact that a change can have in the variant site (16).

VARIANT also reports the information available for already described variants such as Single Nucleotide Polymorphisms (SNPs). This information is collected from different databases: dbSNP (17), Ensembl (12) and 1000 genomes (18). Annotated SNPs are taken from Ensembl's Application Programming Interface (API) which integrates distinct databases such as: HGMD-PUBLIC (19), NHGRI GWAS catalog (20), OMIM (http://omim.org/) and Open Access GWAS Database (21). We also include pathologic mutations collected from COSMIC (http://www.sanger.ac.uk/genetics/CGP/cosmic/) and UniProt (22).

The dilemma on where to place the heavy data

In the process of annotation of variants, there are two heavy data: the own VCF data and the already big and fast growing databases used for the annotation. The use of Web interfaces or Web services has been almost discarded given the difficulties of transferring variation files through the internet in favour of the local run. Unfortunately, local run requires of the local installation of the database, which present an amazing growing rate. Following a philosophy similar to Google, here we choose an innovative solution: heavy data are not moved through the web. The VCF is processed locally with a CLI client that, via RESTful Web Services send batches of queries in parallel to the remote Web service (alternatively the Web service can be used by another local script using the available API). This client has been implemented to send only the information needed to obtain the consequence type of variants, by doing this, data transfer is minimized. Then an optimized Java server program queries the database on the server side and returns the response to the client. This process mimics a local process minimizing data transfer. The resulting query process is efficient and very fast and returns the annotation of >10 000 variants per minute. Figure 1 shows a schema of the client-server architecture. A main collateral advantage of this scenario is that no installation or update of databases are needed. The database with the required information is always up-to-date in the remote server. Our group has maintained updated databases of functional effect of variants for different SNP-related projects (23–25).

Other technical features

VARIANT is a Java application that can run in any platform. The database is queried in an optimized way by RESTful Web Services, either directly or invoked through a CLI program. There is also the possibility of using VARIANT as a Google Chrome Extension, having information of any variant with a mouse click. A RESTful Web Service API to calculate consequence types has been implemented in Java to make accessible the information contained in the database with simple calls. For example:

http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/genomic/variant/13:32906982:T/consequence_type

returns the consequence type of a substitution of the reference base by a T in the position 32906982 of the chromosome 13.

VARIANT comes with a VCF Genome browser developed in HTML5 with Scalable Vector Graphics (SVG) that allows representing the variants found in their genomic context and offers the possibility of visualizing information on the genes, their properties and any other genomic feature around. Figure 2 shows several views that can be displayed by the Genome

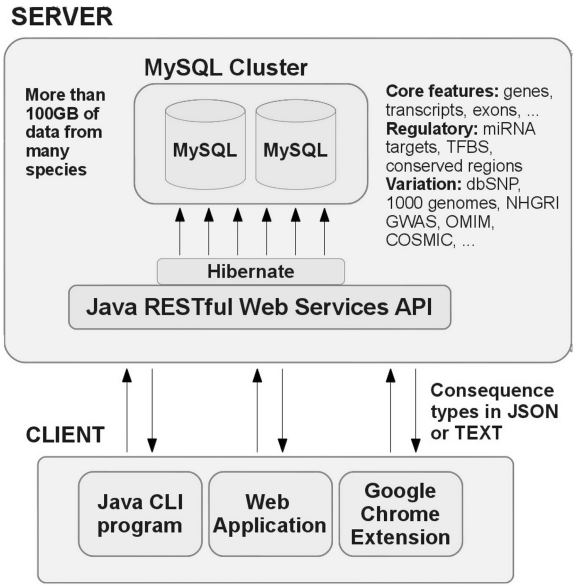

Figure 1. Schema of the client-server architecture of VARIANT. Biological information is stored in a remote MySQL cluster which is accessed through a Java RESTful WEB Services API. To connect to database Hibernate library is used. Data can be retrieved by clients in both text and JSON formats.

Browser. In the upper part a pie chart summarizes the type of variants found and two bar charts represent the distribution of variant across chromosomes and in terms of quality. The central and lower parts show two filters to represent the variants in the genome browser. Using the different filters provided by the tool, different representations of the variants in their genomic context, along with the associated genomic and functional information, can be obtained.

Another interesting feature, first implemented in our program, is that registered users can consult their launched jobs at any time and place, given that sessions and data are stored in our servers.

Other tools

Prediction of the putative functional effect of a mutation is a classic problem already addressed in the context of studies of associations using SNPs (26), and several popular tools have been used for this purpose such as PolyPhen (27) and SIFT(28) or PupaSuite (24). However, the use of exome sequencing has changed the nature of the problem in two respects: (i) contrarily to the case of SNPs, the variants found in NGS studies are, in many cases, unknown and (ii) the number of variants to be annotated is much higher than in the case of a conventional SNP association study. In the last two years, a few tools have been specifically designed to cope with this new challenge in the functional annotation of variants. The first and probably the most popular tools are Annovar (7) and snpEff (http://snpeff.sourceforge.net/). Both are local tools that require the installation of a database, are reasonably fast and provide a succinct, but useful, annotation of the variants (essentially consequence type and some additional information). Recently, other applications more oriented to human genomes like SVA (29), which also offer a convenient Java-based interface, or TREAT (30) have been published. Other applications covering more genomes are also recently available (31).

According to the information given in the respective publications, runtimes are quite similar (approximately one exome, assuming 30–40,000 variants per exome in half hour), except for the snpEff, which claims to make over a million predictions per minute (according to http://snpeff.sourceforge.net/). VARIANT would be in an intermediate place between these runtimes, with, approximately, an exome in less than five minutes (or several hundreds of predictions per second), taking into account that the program also searches regulatory and variation information.

The growing sizes of the databases will be a limiting factor for future local usage. For example, the TREAT (30) database requires 175 GBs, and this data size will increase as more biological information is added and more genomes are sequenced.

The extended use of scripting programming languages, such as Perl or Phyton, used in programs like ANNOVAR, Variant Effect Predictor, NGS-SNP (31) and TREAT can make the programming step easier but at the exchange of immense increases in both runtimes and difficulties for the scalability of future releases.

DISCUSSION

Current SNV annotation tools have different limitations. Most of them only report information on SNPs already present in dbSNP or 1000 genomes, or a few functional features, such as the consequence type or information on diseases or phenotypes. Generally speaking any variant labelled as non-coding or synonymous was filtered out. VARIANT increases the information scope outside the coding regions by including all the available information on regulation, DNA structure, conservation, evolutionary pressures, etc. Regulatory variants constitute a recognized, but still unexplored, cause of pathologies (32). The determination of variants with potential regulatory effect can explain many phenotypes or susceptibilities. As an example of the importance of the regulatory variants, the potential role of CDKN2B in the development of sporadic medullary thyroid carcinoma was confirmed by our group using a functional assay that showed that a variant (the SNP rs7044859) in the promoter region of the gene altered the binding of the transcription factor HNF1 (33).

Another innovative aspect of VARIANT is the client-server architecture that separates physically the database, which is remote, from the local execution. This original solution minimizes the data traffic through internet and frees the user from disk space constraints and the need of cumbersome database updating processes. This client-server process almost mimics a local process minimizing data transfer. The resulting query is extremely fast and returns the annotation of ~10 000 variants per minute.

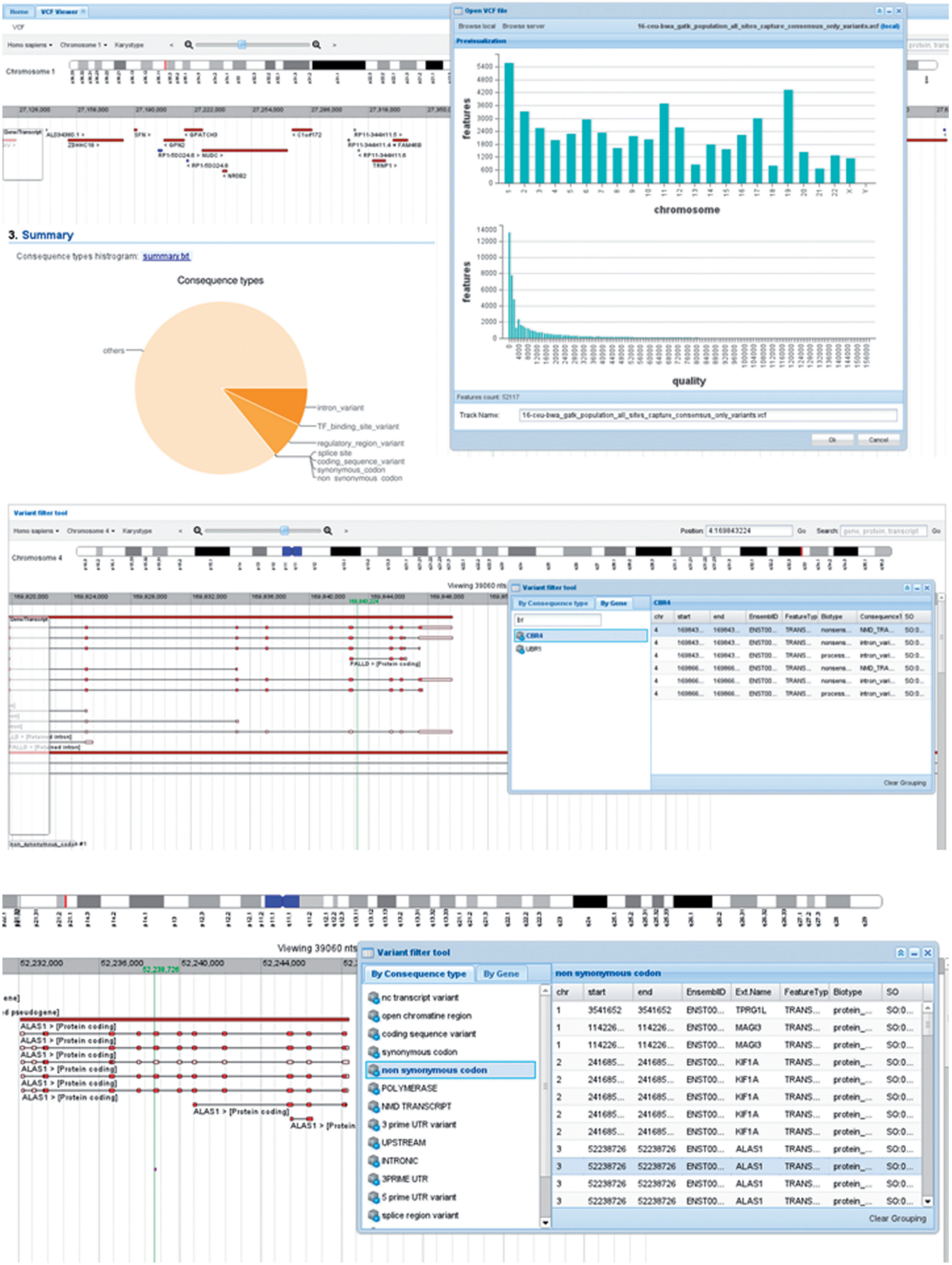

Figure 2. Different representative views that can be displayed by the Genome Browser. In the upper part a pie chart summarizes the type of variants found and two bar charts represent the distribution of variant across chromosomes and in terms of quality. The central part displays the gene filter and the lower part shows the variant-type filter. Using the different filters, different representations of the variants in the genomic context can be obtained.

All these features make VARIANT a comprehensive, fast and innovative tool for the annotation of variants found in exome or genome sequencing experiments.

FUNDING

Funding for open access charge: The Spanish Ministry of Science and Innovation (MICINN) [BIO2011-27069]; the Conselleria de Educacio of the Valencian Community [PROMETEO/2010/001]; National Institute of Bioinformatics (www.inab.org) and the CIBER de Enfermedades Raras (CIBERER), both initiatives of the ISCIII, MICINN; Red Tematica de Investigacion Cooperativa en Cancer (RTICC), ISCIII, MICINN [RD06/0020/1019]; 'Programa Nacional de Proyectos de investigación Aplicada' [I+D+i 2008]; 'Subprograma de actuaciones Científicas y Tecnológicas en Parques Científicos y Tecnológicos' [ACTEPARQ 2009]; FEDER.

Conflict of interest statement. None declared.

REFERENCES

1. Bamshad,M.J., Ng,S.B., Bigham,A.W., Tabor,H.K., Emond,M.J., Nickerson,D.A. and Shendure,J. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.*, **12**, 745–755.
2. Majewski,J., Schwartzentruber,J., Lalonde,E., Montpetit,A. and Jabado,N. (2011) What can exome sequencing do for you? *J. Med. Genet.*, **48**, 580–589.
3. Pop,M. and Salzberg,S.L. (2008) Bioinformatics challenges of new sequencing technology. *Trends Genet.*, **24**, 142–149.
4. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
5. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
6. Sana,M.E., Iascone,M., Marchetti,D., Palatini,J., Galasso,M. and Volinia,S. (2011) GAMES identifies and annotates mutations in next-generation sequencing projects. *Bioinformatics*, **27**, 9–13.
7. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
8. McLaren,W., Pritchard,B., Rios,D., Chen,Y., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.
9. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
10. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
11. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnol.*, **25**, 1251–1255.
12. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.
13. Bryne,J.C., Valen,E., Tang,M.H., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
14. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
15. Altshuler,D.M., Gibbs,R.A., Peltonen,L., Dermitzakis,E., Schaffner,S.F., Yu,F., Bonnen,P.E., de Bakker,P.I., Deloukas,P., Gabriel,S.B. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
16. Capriotti,E., Arbiza,L., Casadio,R., Dopazo,J., Dopazo,H. and Marti-Renom,M.A. (2008) Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. *Hum. Mutat.*, **29**, 198–204.
17. Bhagwat,M. (2010) Searching NCBI's dbSNP database. *Curr. Protoc. Bioinformatics*, Chapter 1, Unit 1.19.
18. The_1000_Genomes_Project_Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
19. Cooper,D.N., Stenson,P.D. and Chuzhanova,N.A. (2006) The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr. Protoc. Bioinformatics*, Chapter 1, Unit 1.13.
20. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
21. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.
22. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
23. Conde,L., Vaquerizas,J.M., Santoyo,J., Al-Shahrour,F., Ruiz-Llorente,S., Robledo,M. and Dopazo,J. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.*, **32**, W242–W248.
24. Conde,L., Vaquerizas,J.M., Dopazo,H., Arbiza,L., Reumers,J., Rousseau,F., Schymkowitz,J. and Dopazo,J. (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res.*, **34**, W621–W625.
25. Saar,K., Beck,A., Bihoreau,M.T., Birney,E., Brocklebank,D., Chen,Y., Cuppen,E., Demonchy,S., Dopazo,J., Flicek,P. *et al.* (2008) SNP and haplotype mapping for genetic analysis in the rat. *Nat. Genet.*, **40**, 560–566.
26. Karchin,R. (2009) Next generation tools for the annotation of human SNPs. *Brief Bioinform.*, **10**, 35–52.
27. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
28. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
29. Ge,D., Ruzzo,E.K., Shianna,K.V., He,M., Pelak,K., Heinzen,E.L., Need,A.C., Cirulli,E.T., Maia,J.M., Dickson,S.P. *et al.* (2011) SVA: software for annotating and visualizing sequenced human genomes. *Bioinformatics*, **27**, 1998–2000.
30. Asmann,Y.W., Middha,S., Hossain,A., Baheti,S., Li,Y., Chai,H.S., Sun,Z., Duffy,P.H., Hadad,A.A., Nair,A. *et al.* (2012) TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. *Bioinformatics*, **28**, 277–278.
31. Grant,J.R., Arantes,A.S., Liao,X. and Stothard,P. (2011) In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics*, **27**, 2300–2301.
32. Hudson,T.J. (2003) Wanted: regulatory SNPs. *Nat. Genet.*, **33**, 439–440.
33. Ruiz-Llorente,S., Montero-Conde,C., Milne,R.L., Moya,C.M., Cebrian,A., Leton,R., Cascon,A., Mercadillo,F., Landa,I., Borrego,S. *et al.* (2007) Association study of 69 genes in the ret pathway identifies low-penetrance loci in sporadic medullary thyroid carcinoma. *Cancer Res.*, **67**, 9561–9567.